Estudio de un sistema de reconocimiento biométrico mediante firma manuscrita online basado en SVM usando Análisis Formal de Conceptos

Aitor Mendaza Ormaza[1], Oscar Miguel Hurtado[1], Raúl Sánchez Reillo[1],
Francisco Valverde Albacete[2], Carmen Peláez Moreno[2],
[1] Grupo Universitario de Tecnologías de Identificación (GUTI), Tecnología Electrónica,
Universidad Carlos III de Madrid, Avda. Universidad 30,
E-28911 Leganés (Madrid), Spain
{amendaza, omiguel, rsreillo}@ing.uc3m.es
[2] Departamento de Teoría de la Señal y Comunicaciones,
Universidad Carlos III de Madrid, Avda. Universidad 30,
E-28911 Leganés (Madrid), Spain
{fva, carmen}@tsc.uc3m.es

Abstract. En el presente artículo se pretende estudiar las prestaciones de un sistema de reconocimiento biométrico mediante firma manuscrita usando la teoría de Análisis Formal de Conceptos (FCA). Se usará la modalidad online de la firma manuscrita, con un algoritmo basado en Máquinas de Vectores Soporte (SVM). Para analizar el desempeño del sistema se realizará un estudio de su matriz de confusión usando el Análisis de Conceptos Formales, y se procederá a extraer conclusiones sobre el sistema.

Keywords: Biometría, firma manuscrita, online, Support Vector Machine, SVM, Formal Concept Análisis, confusion matrices.

1 Introducción

La firma manuscrita es la forma más difundida para la acreditación personal en la vida cotidiana. Se usa en el comercio y en transacciones bancarias, pagos mediante tarjetas de crédito y, en general, toda clase de documentos legales. Por lo tanto, dentro de todas las modalidades biométricas, la firma es probablemente la que más aceptación tiene en el día a día, ya que además no necesita métodos invasivos de medida. Por otro lado, al contrario que otras modalidades biométricas como el iris o la huella dactilar, la firma es una característica de comportamiento de los individuos, y por lo tanto, se considera más débil frente al fraude.

El presente artículo se ha organizado de la siguiente forma: en la sección 2 se presenta un breve estado del arte. Se introduce brevemente el reconocimiento biométrico de firmantes en la modalidad on-line. Posteriormente se hace una breve introducción a las Máquinas de Vectores Soporte SVM, para pasar a realizar una breve introducción sobre el Análisis Formal de Conceptos. Por último se introduce la base de datos usada para la evaluación del sistema biométrico. En la siguiente sección

(3) se procede a exponer los resultados obtenidos mediante el análisis formal de conceptos, empleando para ello la exploración de retículos. Por último, en la sección 4 se incluyen una serie de líneas futuras, así como unas breves conclusiones obtenidas del análisis del sistema mediante el FCA.

2 Estado del Arte

A continuación se introduce un breve estado del arte sobre las distintas técnicas que se han empleado en la preparación del presente artículo.

2.1 Reconocimiento biométrico mediante firma manuscrita

La verificación de firma es el proceso mediante el cual, dada una firma que pertenece a un usuario, una decisión se toma sobre si dicha firma ha sido hecha por ese usuario, una firma genuina, o ha sido realizada por otro usuario, una firma falsificada. Típicamente, las firmas falsificadas se clasifican en tres grupos: (1) aleatoria, (2) simple y (3) experta. Las falsificaciones aleatorias se realizan sin ningún conocimiento sobre las firma del usuario o de su nombre. Las falsificaciones simples se realizan sabiendo el nombre del usuario pero sin ningún conocimiento sobre su firma. Las falsificaciones expertas se realizan con un conocimiento completo sobre el nombre y la firma del usuario.

Los diferentes métodos para la verificación de firma pueden dividirse en dos grupos principales: off-line (estáticos) y on-line (dinámicos). Las técnicas off-line se basan en procesar una imagen digitalizada en escala de grises de la firma manuscrita en un papel [1]. Por el contrario, las técnicas on-line tienen en cuenta las características dinámicas de la firma, tales como la presión ejercida, las inclinaciones, posiciones o la velocidad del stylus. Todas las señales proveen, no sólo de información de la firma, sino también información sobre el acto propio de firmar, que se considera relacionado con el usuario específico (una característica de comportamiento). Debido a esto, y a la mayor cantidad de información recogida, los sistemas de verificación automáticos on-line obtienen una mayor fiabilidad que los sistemas off-line. Las diferentes técnicas de procesado de señal de otros autores varían desde los algoritmos genéticos [2] a las transformadas wavelets [3]. Otros autores usan Modelos Ocultos de Markov (HMM) [4], [5] y [6], lógica difusa [7], distancias euclídeas [8], redes neuronales o Mezcla de Gaussianas (GMMs). En el presente artículo se han usado Máquinas de Vectores Soporte (SVM). Las máquinas de vectores soporte fueron introducidas por V. Vapnik y C. Cortes [9] para problemas de clasificación binarios.

Para el problema bajo estudio, firmas on-line, el sensor que captura los datos suele ser una tableta gráfica. Estos dispositivos proveen, para una firma dada, la evolución temporal de 5 funciones: las coordenadas X, Y, la presión ejercida, y la orientación del bolígrafo (inclinación y azimut). Los valores de estas funciones están muestreadas a un periodo dado. En la Figura 1 se representa la evolución temporal de estas cinco funciones. Una vez se ha obtenido la información de la firma, se realiza un preprocesado. Se aplica un filtro pasobajo para eliminar el ruido introducido durante

la adquisición. A continuación, se aplica una normalización a estas funciones temporales. Se extrae un conjunto de características de las muestras de la firma, tal como se explica en [10].

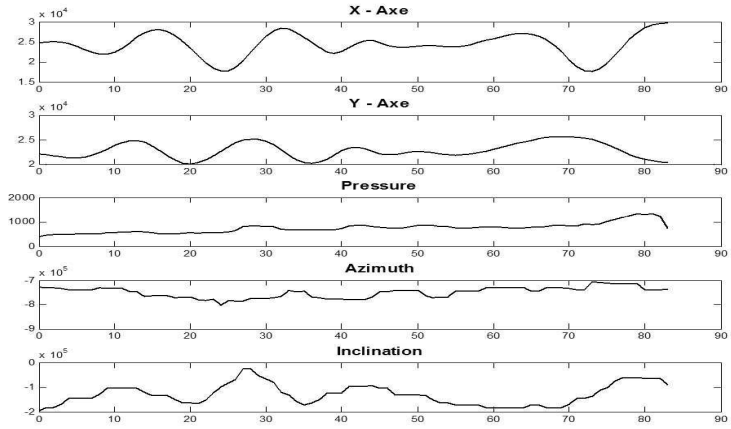

Fig. 1. Ejemplos de señales capturadas mediante una tableta.

Mediante este conjunto de características se debe poder distinguir diferencias entre firmas de diferentes individuos y, al mismo tiempo, reconocer similitudes entre aquellas firmas del mismo individuo. Por último, este conjunto de características se compara con un patrón previamente reclutado y almacenado. Esta comparación nos da una similitud s entre las dos muestras, que se usa para tomar una decisión basada en un umbral U. Si s es mayor que U, el sistema decide que las dos firmas proceden de un mismo usuario.

En el campo de la biometría, a la hora de realizar verificación, se realiza una comparación uno-a-uno (1:1). Se trata de verificar si el usuario sometido al reconocimiento es quien dice ser. Las características biométricas obtenidas en el proceso de reconocimiento se comparan con los patrones almacenados en la base de datos para hallar una coincidencia.

Con el experimento de verificación se pretende obtener una medida de las prestaciones del algoritmo, obteniendo gráficas en las que se muestran las probabilidades de Falsa Aceptación FA, contra las probabilidades de Falso Rechazo FR. En estas gráficas podemos localizar el punto óptimo para configurar el algoritmo, el Equal Error Rate EER, que es el punto donde se cruzan las gráficas de FR con FA. Mediante el experimento de identificación se procederá a extraer la matriz de confusión que se usará en el Análisis Formal de Conceptos.

2.2 Máquinas de Vectores Soporte SVM

Las Máquinas de Vectores Soporte (SVM) es un método de aprendizaje introducido por V. Vapnik [9] y [11], para problemas de clasificación binarios. La máquina propone proyectar los datos de entrada, mediante una proyección que generalmente es no lineal, a un espacio de características F, de muy alta dimensión. En este espacio

proyectado se toma una superficie de decisión (un hiper-plano) que maximiza la distancia de ambas clases al hiper-plano y separa el mayor número posible de puntos perteneciente a la misma clase en el mismo lado (margen máximo entre los vectores de las dos clases). Así pues, el error de clasificación de ambos datos en el conjunto de entrenamiento y de test, se minimiza.

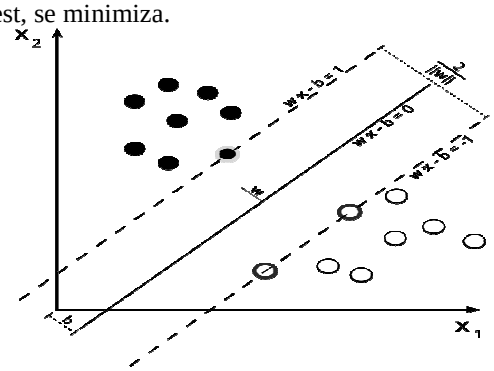

Fig. 2. Hyperplano de Máximo- Margen y los márgenes de un SVM entrenado con muestras de dos clases.

Los algoritmos básicos de las SVMs usan umbrales lineales. Pero mediante un simple cambio de la función (kernel) K(u,v) del algoritmo, pueden usarse las SVMs para que aprendan otros tipos de umbrales, tales como:

- Polinómicos (homogeneos): $k(x_i, x_j) = (x_i \cdot x_j)^d$
- Polinómicos (inhomogeneos):

$$k(x_i, x_j) = (x_i \cdot x_j + 1)^d$$

- Funciones de Base Radial (RBF):

$$k(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right), \text{ for } \gamma > 0$$

- Redes neuronales sigmoides de N capas.

Tal como se ha comentado previamente, las SVMs son clasificadores binarios. Para el problema multiclase de reconocimiento de firmas, se ha reducido el problema a múltiples problemas binarios. Existen dos métodos para construir dichos clasificadores binarios: i) uno-contra-todos (one-versus-all), donde el clasificador distingue entre una de las clases y el resto como segunda clase; ii) uno-contra-uno (one-versus-one), donde el clasificador distingue entre cada par de clases. Se ha usado el enfoque de uno-contra-todos debido a que en un sistema de reconocimiento de firmas real, el número de usuarios es muy grande, y puede crecer rápidamente. Para el entrenamiento se toman N firmas genuinas para una clase, y M falsificaciones expertas para la otra clase, teniendo un conjunto (set) de (N,M) firmas. Con este juego de firmas, se entrena una SVM y se obtiene un modelo para cada usuario, que será usado posteriormente durante la fase de test para determinar si una determinada firma pertenece a un usuario.

2.3 Análisis Formal de Conceptos FCA

El Análisis Formal de Conceptos es una de las formas principales de derivar una ontología a partir de una colección de objetos y propiedades. El término fue introducido por Rudolf Wille en 1984. Se ha aplicado en conceptos como la teoría de órdenes o el estudio de retículos. El Análisis Formal de Conceptos puede describirse tanto como una técnica de aprendizaje máquina no supervisado, o como un método de análisis de datos. En el FCA se definen una serie de *estímulos* de entrada, y su correspondiente salida, llamadas *respuestas*. A los *estímulos* se les suele llamar también *objetos*, mientras que a las *respuestas* se las conoce con el nombre de *atributos* o *propiedades*.

Introducimos ahora la idea de *contexto* y *concepto*. Un *contexto formal* (V_A, V_B, I_{CM}) es un conjunto de *estímulos* u *objetos* (V_A), un conjunto de *respuestas* o *atributos* (V_B), y una función que indica la relación entre los *estímulos* y las *respuestas* (I_{CM}). En el problema bajo estudio, dicha relación se dará en forma de una *Matriz de Confusión CM*. Un *concepto formal* es un subconjunto de *estímulos* totalmente relacionado (o confundido en nuestro caso) con otro subconjunto de *respuestas*. En este caso, al subconjunto de *estímulos* se le llama *extensión* (*extent*), y al subconjunto de *respuestas* se le conoce como *intención* (*intent*).

Los *conceptos* definidos arriba pueden ordenarse parcialmente mediante inclusión. Si se cumple la relación siguiente: $(A_i, B_i) \leq (A_j, B_j) \Leftrightarrow A_i \subseteq A_j \Leftrightarrow B_i \supseteq B_j$, podemos decir que el primer concepto Ci es más específico (menos general) que el segundo concepto Cj. Al mayor de los límites inferiores se le conoce como *meet*. Al menor de los límites superiores se le conoce como *join*. Mediante estas operaciones de *meet* y *join*, se satisfacen los axiomas necesarios para definir un retículo. En este caso, dado que el retículo se genera a partir de una *matriz de confusión*, diremos que se trata de un *retículo de confusión*. En general, los *retículos de confusión* se suelen generar a partir de matrices booleanas de confusión. Para el problema bajo estudio, se va a emplear la técnica descrita en [12] para explorar una matriz de confusión no booleanas, así como sus retículos asociados.

Usando el *K-Análisis Formal de Conceptos* K-FCA [13] y [14], se introduce la noción de *grado de incidencia*. Con esta noción, las matrices de confusión usadas pueden ser no booleanas, y podemos decir que un *estímulo* a_i se confunde con una *respuesta* b_j con *grado* (de confusión) λ. Se define además un *umbral de existencia* φ, que se usará para decidir si un *estímulo* y una *respuesta* se consideran que están confundidos. Dicha confusión se produce cuando el *grado* λ es superior al *umbral* φ. Variando el *umbral* φ se realiza una *exploración de retículos*. Para comparar distintos retículos durante la exploración, se usa el *contador de conceptos*, que representa el número de nodos presentes en el correspondiente retículo, y proporciona una medida de la complejidad de la representación reticular.

2.4 Base de Datos usada

El estudio de las características extraídas de una firma y la evaluación experimental del sistema de verificación de firmas on-line se ha llevado a cabo gracias a la Base de

Datos de firmas MCyT, que está distribuida de forma pública [15]. Esta Base de Datos consiste en 100 usuarios diferentes. Cada usuario ha generado 25 firmas genuinas, se han producido 25 falsificaciones expertas para cada usuario. Estas falsificaciones expertas han sido producidas por los 5 usuarios siguientes, que han tenido conocimiento de la firma y la han practicado hasta que se han sentido fluidos en la falsificación de la firma. Para capturar la base de datos se ha usado una tableta gráfica Wacom Intuos A6 USB. Esta tableta facilita las siguientes funciones temporales, de forma discreta (también se especifican los rangos de cada función), tal como se muestra en la Figura 1: i) posición en el eje x [0-12700]; ii) posición en el eje y [0-9700]; iii) presión aplicada por el bolígrafo [0-1023]; iv) ángulo azimut [0-3600]; v) ángulo de inclinación [0-900]. También se capturan los movimientos de empezar a escribir con el stylus o bolígrafo, y levantarlo. La frecuencia de muestreo se fijó a 100Hz.

3 Resultados obtenidos

A continuación se van a exponer los resultados obtenidos mediante el FCA y la exploración de retículos, definida previamente. La matriz de confusión obtenida se puede observar en la Figuras 3 (también conocida como representación *heapmap* de la matriz de confusión).

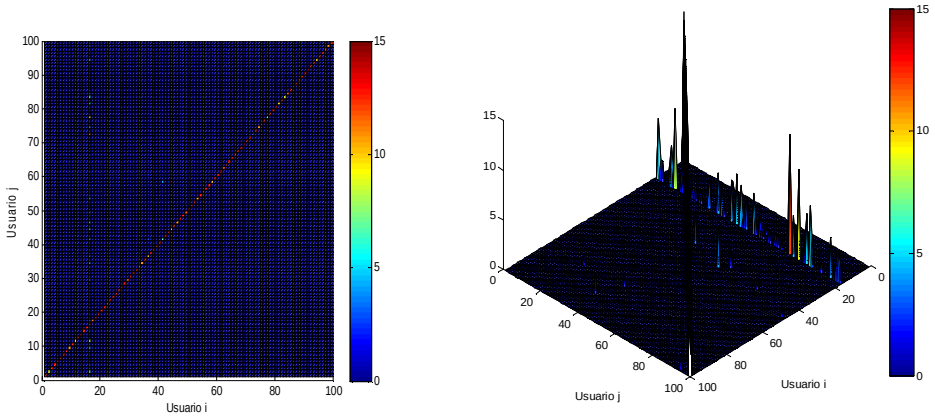

Figura 3. (izq)Representación 2D de la matriz de confusión. (dcha) Representación 3D de la matriz de confusión

3.1 Evolución del contador de Conceptos

Se han tomado una serie de valores para el *umbral* φ, y se ha calculado su retículo correspondiente. En la Tabla 1, así como en la Figura 4, podemos ver el resumen de

los valores de los umbrales elegidos, así como del número de conceptos calculados en el retículo para cada umbral.

Umbral φ	#(conceptos)	Umbral φ	#(conceptos)
-0.613532	106	3.321928	107
0.386468	106	3.556393	107
0.971431	107	3.736966	107
1.386468	107	4.321928	107
1.708396	107	5.058894	107
1.971431	107	5.866249	107
2.386468	107	6.058894	107
2.556393	107	6.463284	107
2.643856	107	6.491853	107
2.736966	107	6.536941	107
2.836501	107	6.544321	107
2.971431	107	6.550747	107
3.058894	107	6.643856	V
3.293359	107		

Tabla 1. Umbrales usados, así como el número de conceptos calculados.

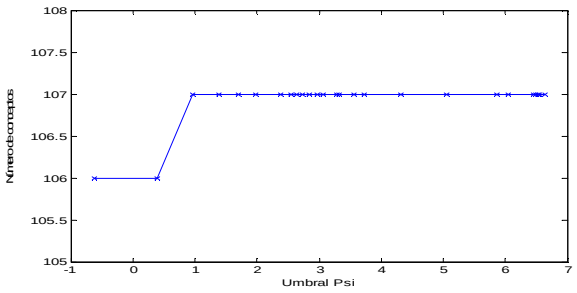

Figura 4. Número de conceptos vs umbrales usados.

Como se puede comprobar, el número de conceptos permanece más o menos constante, entre 106 y 107. De aquí podemos extraer que tenemos una matriz muy homogénea, y bastante buena, a priori. A continuación, vamos a estudiar los retículos generados para extraer más información sobre el clasificador.

3.2 Exploración de retículos

En este apartado vamos a representar varios de los retículos estudiados. Para ello nos apoyaremos del programa Concept Explorer v1.3 [16]. Como se puede comprobar en la Figura 5 y tal como se ha comentado anteriormente, tenemos unos retículos con un gran número de conceptos (106-107). Se pueden observar numerosos estímulos perfectamente clasificados. Estos estímulos son los que tienen como padre el concepto *top* y como hijo el concepto *bottom*. Para que resulte más sencillo el análisis de los retículos, vamos a proceder a podar todos los estímulos perfectamente clasificados

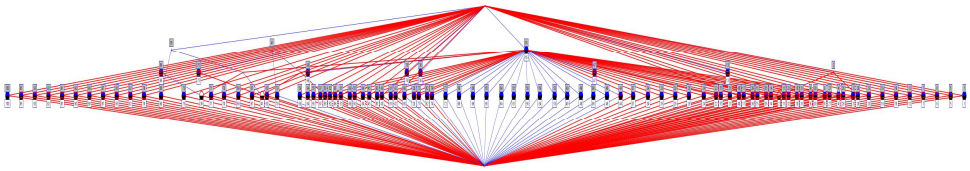

Figura 5: Retículo completo para umbral φ = 3.556393.

Para ilustrar las conclusiones obtenidas, se ha decidido mostrar tan sólo tres de todos los retículos estudiados. Se han usado los umbrales φ = -0.613532, φ = 3.556393 y φ = 6.550747, cuyos retículos podados se pueden ver en las Figuras 6, 7 y 8, respectivamente. Partiendo de estudios previos, cabría esperar que para valores pequeños del umbral tuviésemos un retículo bastante ruidoso, con gran confusión, y difícil de interpretar. Al ir aumentado el valor del umbral, el número de conceptos se reduciría paulatinamente, ofreciendo un retículo mucho más simple, en los que quedarían de manifiesto las confusiones más evidentes.

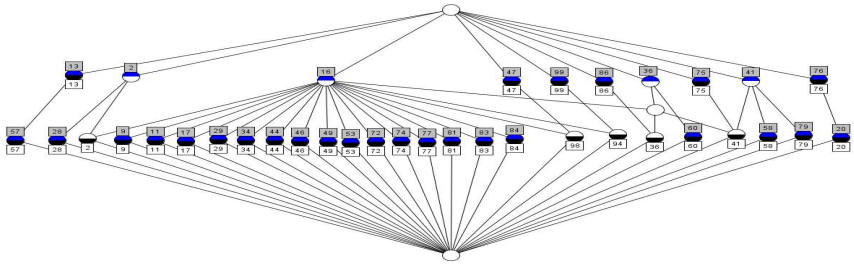

Figura 6: Retículo podado para umbral φ = -0.613532

Sin embargo, al ver las Figuras 8 y 9, y sabiendo que el número de conceptos no se ha reducido con todos los umbrales estudiados, podemos observar que para la matriz de confusión bajo estudio, tal como se ha comentado previamente, es una matriz muy homogénea, que ofrece unas buenas prestaciones. Sin embargo, podemos observar como tenemos un error sistemático, que crece al aumentar el umbral. Para un umbral pequeño (Figura 7), tenemos dos grandes grupos. Un grupo de estímulos, que se confunden sistemáticamente con la respuesta 16, y por otro lado, otro grupo de estímulos, o subretículos paralelos, creando canales virtuales para estos usuarios.

Al aumentar el tamaño del umbral (Figuras 8 y 9) podemos ver como los subretículos paralelos que se podían observar en la Figura 7, van desapareciendo para pasar a formar parte del subretículo que tiene como *meet-irreducible* a la respuesta 16.

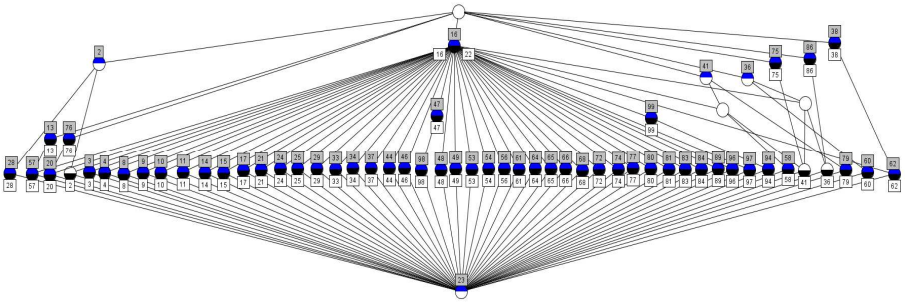

Figura 7: Retículo podado para umbral φ = 3.556393

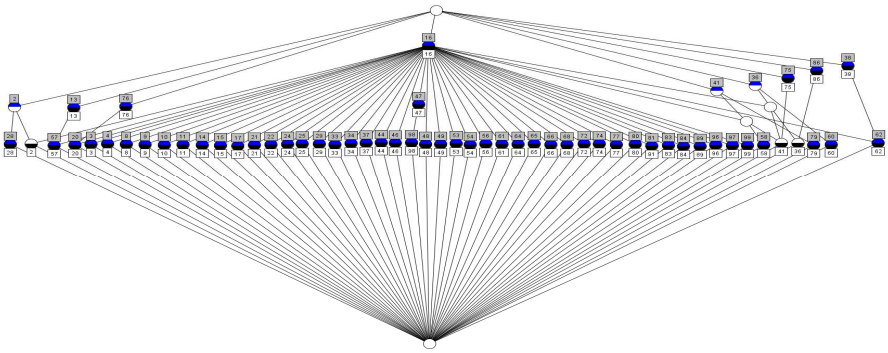

Figura 8: Retículo podado para umbral φ = 6.550747

Queda de manifiesto que el sistema de reconocimiento bajo estudio está cometiendo una serie de errores sistemáticos con el usuario 16.

Existen también ciertas confusiones paralelas al subretículo del usuario 16, pero las confusiones son pocas, y con un grado de confusión menor. Además, podemos observar que no tenemos ningún concepto *totalmente confundido*, esto es, que el sistema es incapaz de distinguir entre los elementos del concepto.

4 Conclusiones y líneas futuras

Gracias al FCA hemos podido ver como se comporta el sistema de identificación bajo estudio. Hemos visto que el sistema ofrece unos buenos resultados, con un índice de confusión bastante bajo. Usando los retículos creados a partir de la matriz de confusión, hemos podido detectar un usuario con el cual se confunden sistemática. La detección de este problema ha sido gracias al uso del FCA. Este tipo de análisis no se suele usar en los sistemas de identificación biométrica. Al usarlo hemos podido ver como existen ciertos fallos.

Como trabajo futuro, y gracias al análisis realizado, se pretende seguir investigando las causas del error sistemático del usuario 16, así como analizar las posibles debilidades y fortalezas del sistema. Para ello, se estudiará de forma gráfica

las firmas correspondientes a los usuarios que se confunden con el usuario 16. Otra posible opción es intentar ajustar el algoritmo de forma especial para el usuario 16.

References

1. Sabourin, R., Plamondon, R., Lorette, G.: Off-line Identification with Handwritten Signature Images: Survey and Perspectives. In: IAPR workshop on Syntactic and Structural Pattern Recognition, AT&T Murray Hill, New Jersey, pp 377--391, (June 1990)
2. Galbally, J., Fierrez, J., Freire, M.R., Ortega, J.: Feature Selection Based on Genetic Algorithms for On-Line Signature Verification. In: IEEE Workshop on Automatic Identification Advanced Technologies, pp. 198-203 (2007)
3. Letjam, D., George, S.: On-line handwritten signature verification using wavelets and back-propagation neural networks. In: Proc. Of ICDAR'01, Seattle, pp. 596-598 (2001)
4. Yang, L., Widjaja, B.K., Prasad, R.: Application of hidden Markov models for signature verification. In: Pattern Recognition, vol. 28, No. 2, pp. 161-170 (1995)
5. Baum, L.E.: An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. In: Inequalities, vol.3, pp 1-8 (1972)
6. Baum, L.E., Egon, J.A.: An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology. Bull. Amer. Meteorol. Soc., vol 73, pp 360-363 (1967)
7. Mingming, Ma., Wijesona, W.S.: Automatic on-line signature verification based on multiple models. In: Proceedings of the IEEE/IAFE/INFORMS 2000 Conference on Computational Intelligence for Financial Engineering, (CIFEr), pp. 30-33 (2000)
8. Diamuro, G., Impedovo, S., Pirlo, G.: A stroke-oriented approach to signature verification. In: From Pixels to Features III - Frontiers in Handwriting Recognition, S. Impedovo and J.C. Simon eds, Elsevier Publ., pp 371-384 (1992)
9. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning. vol. 20, pp. :273-297, (1995)
10. Mendaza-Ormaza, A., Miguel-Hurtado, O.: On-line Signature Biometrics using Support Vector Machine. In: BIOSIG, volume 155 of LNI, page 185-188. GI, (2009)
11. Vapnik, V.: The Nature of Statistical Learning Theory. Springer (1995)
12. Peláez-Moreno, C., García-Moral, A. I., Valverde-Albacetece, F. J.: Analyzing phonetic confusions using Formal Concept Análisis. Signal Theory and Communications Department, EPS-Universidad Carlos III de Madrid, Leganés-28911, Spain. January 25 (2010)
13. Valverde-Albacete, F. J., Peláez-Moreno, C.: Towards a generalisation of Formal Concept Analysis for data mining purposes. In Proceedings of the International Conference on Formal Concept Analysis, ICFCA06, (Dresden, Germany), edited by R. Missaoui and J. Schmid, volume 3874 of Lecture Notes on Artificial Intelligence, 161–176 (Springer, Berlin, Heidelberg) (2006)
14. Valverde-Albacete, F. J., Peláez-Moreno, C.: Galois connections between semimodules and applications in data mining. In: Proceedings of the 5th International Conference on Formal Concept Analysis, ICFCA 2007, Clermont- Ferrand, France, edited by S. Kusnetzov and S. Schmidt, volume 4390 of Lecture Notes on Artificial Intelligence, 181–196 (Springer, Berlin, Heidelberg) (2007)
15. Ortega-Garcia, J., Fierrez-Aguilar, J., Simon, D., Gonzalez, J., Faundez-Zanuy, M., Espinosa, V., Satue, A., Hernaez, I., Igarza, J.-J., Vivaracho, C., Escudero, D., Moro, Q.I.: MCYT baseline corpus: a bimodal biometric database. In: IEEE Proc.-Vis. Image Signal Process., Vol. 150, No. 6, (2003)
16. Concept Explorer, http://sourceforge.net/projects/conexp/